# ADVANCED PROTEOMIC CHARACTERIZATION OF THE 26S PROTEASOME IN ARABIDOPSIS REVEALS INSIGHTS INTO COMPOSITION AND ASSEMBLY

by

David C. Gemperline

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Genetics)

at the

UNIVERSITY OF WISCONSIN–MADISON

2016

Date of final oral examination: TBD

The dissertation is approved by the following members of the Final Oral Committee:
Richard D. Vierstra, Professor, Genetics
Richard Amasino, Professor, Genetics
Josh Coon, Professor, Chemistry
Donna Fernandez, Professor, Botany
Patrick Masson, Professor, Genetics

*To Erin, my wife.*

# ACKNOWLEDGMENTS

*Science doesn't purvey absolute truth. Science is a mechanism, a way of trying to improve your knowledge of nature. It's a system for testing your thoughts against the universe and seeing whether they match. This works not just for the ordinary aspects of science, but for all of life.*

— Isaac Asimov (1988)

Acknowledgements go here.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# LIST OF ABBREVIATIONS AND ACRONYMS

| | |
|---|---|
| API | Application programming interface |
| BCA | Bicinchoninic acid protein assay |
| BLAST | Basic local alignment search tool |
| C# | C sharp, a programming language |
| Da | Dalton, the atomic mass unit |
| DNA | Deoxyribonucleic acid |
| DTT | Dithiothreitol |
| ESI | Electrospray ionization |
| E-value | Expectation value |
| FASTA | A format for storing protein sequences |
| FDR | False discovery rate |
| GUI | Graphical user interface |
| HCD | Higher-energy collisional dissociation |
| HPLC | High-performance liquid chromatography |
| LC | Liquid chromatography |
| *m* | Mass |
| min | Minute |
| MS | Mass spectrometry |

| | |
|---|---|
| $MS^1$ | Survey mass analysis |
| MS/MS | Tandem mass spectrometry |
| NCE | Normalized collision energy |
| nLC | Nanoflow liquid chromatography |
| ppm | Part per million |
| PSM | Peptide-spectrum match |
| PTM | Post-translational modification |
| s | Second |
| SILAC | Stable isotope labeling by amino acids in cell culture |
| S/N | Signal-to-noise ratio |
| TMT | Tandem mass tag |

# ADVANCED PROTEOMIC CHARACTERIZATION OF THE 26S PROTEASOME IN ARABIDOPSIS REVEALS INSIGHTS INTO COMPOSITION AND ASSEMBLY

David C. Gemperline

Under the supervision of Professor Richard D. Vierstra

At the University of Wisconsin-Madison

**Abstract**

The 26S proteasome is the central proteolytic effector in the ubiquitin system that is responsible for degrading numerous regulators following their selective ubiquitylation. While much is known about the construction of the yeast and mammal particles, little is known about the pathways used to assemble the plant particle. One challenge is that the known yeast chaperones appear sufficiently diverged to preclude high-confidence identification of their plant counterparts by genomic searches. Here, we used in-depth mass spectrometric analysis of Arabidopsis 26S proteasomes, which were affinity purified from seedlings under conditions that promote the accumulation of assembly intermediates, to identify a large collection of interacting proteins that associate with either the core protease (CP) or regulatory particle (RP). Sequence comparisons, Y2H and BiFC studies revealed

that some are likely assembly chaperones, with several CP factors harboring the signature C-terminal HbYX motif that allows their association with the α-subunit ring. Several of the RP-specific factors appear to be orthologs of the chaperones Nas2, Nas6, Hsm3 and Ecm29. Whereas yeast assembles only a single particle type, mammals can assemble alternate proteasomes by replacing individual subunits with distinct isoforms (e.g., immunoproteasomes). In plants, most 26S proteasome subunits are encoded by paralogous genes with sufficient divergence to suggest that plants also accumulate a collection of particles. However, proteomic analysis of proteasomes selectively enriched using paralog-specific tags strongly imply that although plants possess this genetic diversity, the incorporation of these paralogs appears random, and is mainly influenced by the differential expression of the corresponding genes. Taken together, these proteomic studies provide the first insights into plant proteasome assembly and diversity, and identify factors that build the CP and RP subcomplexes and finally the 26S holo-particle.

Richard D. Vierstra

# ABSTRACT

**Abstract**

The 26S proteasome is the central proteolytic effector in the ubiquitin system that is responsible for degrading numerous regulators following their selective ubiquitylation. While much is known about the construction of the yeast and mammal particles, little is known about the pathways used to assemble the plant particle. One challenge is that the known yeast chaperones appear sufficiently diverged to preclude high-confidence identification of their plant counterparts by genomic searches. Here, we used in-depth mass spectrometric analysis of Arabidopsis 26S proteasomes, which were affinity purified from seedlings under conditions that promote the accumulation of assembly intermediates, to identify a large collection of interacting proteins that associate with either the core protease (CP) or regulatory particle (RP). Sequence comparisons, Y2H and BiFC studies revealed that some are likely assembly chaperones, with several CP factors harboring the signature C-terminal HbYX motif that allows their association with the α-subunit ring. Several of the RP-specific factors appear to be orthologs of the chaperones Nas2, Nas6, Hsm3 and Ecm29. Whereas yeast assembles only a single particle type, mammals can assemble alternate proteasomes by replacing individual subunits with distinct isoforms (e.g., immunoproteasomes). In plants, most 26S proteasome

subunits are encoded by paralogous genes with sufficient divergence to suggest that plants also accumulate a collection of particles. However, proteomic analysis of proteasomes selectively enriched using paralog-specific tags strongly imply that although plants possess this genetic diversity, the incorporation of these paralogs appears random, and is mainly influenced by the differential expression of the corresponding genes. Taken together, these proteomic studies provide the first insights into plant proteasome assembly and diversity, and identify factors that build the CP and RP subcomplexes and finally the 26S holo-particle.

# Chapter 1

# THE UBIQUITIN 26S PROTEASOME SYSTEM

## 1.1. Ubiquitin Conjugating Machinery

### 1.1.1. E1s

### 1.1.2. E2s

### 1.1.3. E3s

### 1.1.4. DUBS

## 1.2. The 26S Proteasome

### 1.2.1. The 20S Core Protease

### 1.2.2. The 19S Regulatory Particle

#### 1.2.2.1. Regulatory Particle Base

#### 1.2.2.2. Regulatory Particle Lid

### 1.3. Proteasome Expression

### 1.4. Proteasome Assembly

### 1.5. Proteasome Post-Translational Modification

### 1.6. Proteasome Degredation

### 1.7. Proteasome Interacting Proteins

# Chapter 2

# MORPHEUS SPECTRAL COUNTER: A COMPUTATIONAL TOOL FOR LABEL-FREE QUANTITATIVE MASS SPECTROMETRY USING THE MORPHEUS SEARCH ENGINE

## 2.1. Summary

Label-free quantitative MS based on the Normalized Spectral Abundance Factor (NSAF) has emerged as a straightforward and robust method to determine the relative abundance of individual proteins within complex mixtures. Here, we present Morpheus Spectral Counter (MSpC) as the first computational tool that directly calculates NSAF values from output obtained from Morpheus, a fast, open-source, peptide-MS/MS matching engine compatible with high-resolution accurate-mass instruments. NSAF has distinct advantages over other MS-based quantification methods, including a higher dynamic range as compared to isobaric tags, no requirement to align and re-extract MS1 peaks, and increased speed. MSpC features an easy to use graphic user interface that additionally calculates both distributed and unique NSAF values to permit analyses of both protein families and isoforms/proteoforms. MSpC determinations of protein concentration were linear over

several orders of magnitude based on the analysis of several high-mass accuracy datasets either obtained from PRIDE or generated with total cell extracts spiked with purified Arabidopsis 20S proteasomes. The MSpC software was developed in C# and is open sourced under a permissive license with the code made available at `http://dcgemperline.github.io/Morpheus_SpC/`.

### 2.2. Main Text

Quantification of individual polypeptides within complex mixtures by MS is an extremely useful tool to understand proteomic changes in organisms during growth and development, and after environmental perturbation (Wong and Cagney, 2010). While a number of MS/MS strategies have been developed to measure protein abundance, including Stable Isotope Labeling by Amino Acids in Cell Culture (SILAC), labeling with isobaric tags, and Absolute Quantification of proteins (AQUA) (Gerber et al., 2003; Ong et al., 2002; Ross et al., 2004; Thompson et al., 2003), label-free quantification (LFQ) have become increasingly popular given their simplicity and low cost (Wong and Cagney, 2010; Zhang et al., 2006). One LFQ strategy infers abundance from the number of observed peptide spectra matches (PSMs). For these PSM-based approaches, changes in protein abundance can be generated artifactually when total PSMs differ among samples and because longer proteins

tend to produce more raw counts. For these reasons normalizing for both protein length and total PSMs is paramount. While this adjustment can be made in a number of ways; one of the most straight forward methods is to use Normalized Spectral Abundance Factor (NSAF), a length- and count-normalized measure for each protein (Zybailov et al., 2006). Further improvements to the NSAF algorithm have been made by accounting for shared peptides in distributed NSAF (dNSAF), which distributes common PSMs among a family of isoforms/proteoforms based on the number of distinct PSMs observed for each isoform/proteoform, and unique NSAF (uNSAF), which ignores shared PSMs and only assigns distinct PSMs to each specific isoform/proteoform (Zhang et al., 2010).

The Morpheus MS search engine was recently designed for high-resolution, accurate-mass data obtained from Orbitrap-based instruments to provide faster matching of spectra to peptides (Wenger and Coon, 2013). Unfortunately, no downstream automated tools are available to facilitate LFQ analysis, which can be quite challenging, if not impossible, to complete manually when accounting for shared peptides. To overcome this bottleneck, we developed Morpheus Spectral Counter (MSpC) as the first LFQ computational tool that integrates directly with Morpheus to calculate NSAF, dNSAF, uNSAF, and corrected PSM (Fermin et al., 2011) values in complex protein samples. MSpC is fully automated, and only requires a Morpheus

search summary file (summary.tsv) as input. The user interface (Supplemental Figure 1A) allows one to select the summary file and displays the raw MS/MS files that will be analyzed by MSpC. Some important features of MSpC are its ability to handle fractionation experiments as input, and the ability to whitelist proteins of interest in the output by specifying a csv file (see Tutorial). Options exist to specify global PSM and protein group FDR rates (thus avoiding increased FDRs when one analyzes many experiments at once), to output NSAF, dNSAF, and uNSAF values, to require a minimum number of unique peptides to quantify a protein, and to specify an output directory. A progress bar indicates completion of the analysis by MSpC. To validate the accuracy of MSpC, we analyzed two MS/MS datasets available in PRIDE that were previously generated by high-energy collision-induced dissociation using Thermo Q-Exactive Orbitrap instruments. Here, Xenopus egg (top, Figure 1) and embryo (bottom, Figure 1) extracts were spiked at a 4:1 ratio with the Universal Proteome Standard 2 (UPS2), a mix of 48 purified proteins at defined molar ratios of 0.5, 5, 50, 500, 5000, and 50,000, with each ratio containing a different set of 8 of the 48 proteins. As shown in Figure 1A, when the Morpheus/MSpC pipeline was used to calculate the average dNSAF value for each UPS2 protein, requiring only a single unique peptide to quantify, strong linear correlations ($R^2$ = 0.886 and 0.823) were obtained across a 1,000 fold change in abundance (50 fmol

to 50,000 fmol). In fact, the $R^2$ values were similar to those obtained by others with PSM-based LFQ methods (Cox et al., 2014; Tu et al., 2014). This linear correlation was further strengthened when the dNSAF values were averaged for all UPS2 proteins within each of the concentration groups, with $R^2$ values of 0.994 and 0.992 for the egg and embryo datasets, respectively (Figure 1B). Notably, the slope of the concentration series was significantly less than unity, showing that NSAF measurements are not appropriate for absolute quantification, which was expected given that NSAF is a relative value.

We also reprocessed the UPS2 dataset using the option of requiring a minimum of two unique peptides for quantification, which should improve stringency. This option provided only a minor improvement in overall linearity for the average UPS2 dNSAF values, but decreased linearity when each UPS2 protein was considered individually and removed some UPS2 proteins at low concentrations (compare Supplemental Figure 2A to Figure 1A). Consequently, caution should be exercised when selecting this option even though it might provide a slight improvement in stringency (see supplemental discussion in Supporting Information). To demonstrate the utility and accuracy of MSpC as applied to our work, we analyzed 20S proteasomes isolated from Arabidopsis thaliana. This particle contains multiple subunits assembled in stoichiometric amounts, with many subunits encoded by

two paralogous genes of sufficient amino acid identity (typically >90% (**?**)) such that discrimination between paralogs can be challenging using LFQ approaches (Book et al., 2010). To simulate changes in 20S proteasome abundance, we added varying amounts of trypsinized proteasomes (0.05 Âµg to 3 Âµg) to a fixed amount of trypsinized E. coli lysate (0.5 Âµg) to generate proteasome/lysate ratios of 0.091, 0.167, 0.333, 0.500, 0.667, 0.750 0.800, 0.857. The digests were then subjected to MS/MS and the dNSAF value for each subunit along with the uNSAF value for individual isoforms were calculated by the Morpheus/MSpC pipeline (see Supplemental Methods). The data from this experiment are deposited in PRIDE with ID PXD003002. As shown in Figure 2, MSpC provided an excellent determination for the overall abundance of 20S proteasomes within a complex mixture, along with a good reflection of the abundance of individual subunits and their isoforms. When the dNSAF values for all subunits for the Arabidopsis 20S proteasome including their isoforms (representing 14 distinct subunits, 10 of which exist as isoform pairs) were summed, a very close approximation of the dNSAF/actual abundance was obtained (slope=0.875) with a very strong linear correlation ($R^2 = 0.99$) over a 10-fold range in protein abundance. When each 20S proteasome subunit was analyzed individually, a strong linear response was also obtained ($R^2 > 0.90$) for a majority of subunits (Figure 2C and Supplemental Table 1). For example, reasonably accurate

concentration plots were obtained for the PAF (α6) and PBD (β4) subunits that are encoded by the PAF1/2 and PBF1/2 gene pairs, and for the PAG (α7) and PBF (β6) subunits that are encoded by single PAG1 and PBF1 genes ($R^2$ from 0.94 to 0.99). Even when we calculated uNSAF values for individual isoforms added to the E. coli lysate, strong linear responses were obtained (e.g., the PAF1/PAF2 and PBD1/PBD2 pairs) with robust correlations ($R^2$ from 0.89 to 0.95) (Figure 2D). Taken together, MSpC worked well for relative LFQ analysis of a multi-subunit complex and its individual subunits and isoforms within a complex proteomic mixture.

The Morpheus/MSpC pipeline also allowed us to calculate the respective incorporation of each paralog in the complex (see Supplemental Methods). As shown in Figure 2E, these estimated/expected occupancies were close to unity for most subunits within both the α and β rings of the 20S proteasome. The only strong deviation was for PBD1/2 (β4), which had a greater dNSAF value relative to other β subunits across the experiments analyzed (Supplemental Table 1). The calculations for uNSAF values also estimated the relative proportion of each isoform within the complex for those subunits expressed from paralogous genes. The data obtained are similar to prior studies of the complex involving quantitative top-down proteomic analysis of purified proteasome samples using ultra violet-intrinsic fluorescence to quantify tyrosine-containing subunits (Russell et al., 2013). However, our MSpC

analysis provided a more complete picture as several subunit isoforms were difficult to quantify by fluorescence either because they lacked tryosine, or because their fluorescence peaks overlapped with those of other subunits/isoforms. Notably , the protein isoform ratios measured here agree well with the expression ratios for the paralogous genes (Book et al., 2010), suggesting that the protein isoform abundance generally reflects the relative transcriptional activity of the gene pair. We consistently estimated slightly more α ring subunits (PAA-PAG) versus β ring subunits (PBA-PBG) in the final MSpC calculations (Figure 2E). This deviation could represent enhanced detection of α ring versus β ring subunits, or more likely that purification via the tagged α ring subunit PAG1 also isolated assembly intermediates comprised of only α ring subunits.

We compared the Morpheus and MSpC pipeline to the next most comparable open source, spectral-count-based LFQ pipeline, The Trans Proteomic Pipeline (TPP) (Deutsch et al., 2010) and ABACUS (Fermin et al., 2011) using our datasets generated with the 20S proteasome/E. coli lysate mixture (Supplemental Table 1 and 2). Morpheus/MSpC slightly outperformed TPP/ABACUS by having a greater overall accuracy (average linearity of 0.88 compared to 0.84), and by having more subunits showing an $R^2$ linear correlation greater than 0.9 (14/23 subunits for MSpC versus and 11/23 for ABACUS). In addition to this modest improvement, we note

that the Morpheus/MSpC pipeline required significantly less intermediary steps, thus accelerating the data analysis. Some of the additional steps in TPP/ABACUS could be automated from the command-line, but it would likely be a challenge for the average user. Importantly, we found that the Morpheus/MSpC pipeline was faster. Timing tests using the proteasome/E. coli spike data generated here showed that the Morpheus/MSpC pipeline was 1.9-fold faster than the TPP/ABACUS pipeline (Figure 3). Such an improvement was expected given that Morpheus completes its searches on average 1.3 to 4.6 times faster than most other search engines available (Wenger and Coon, 2013). Given its simplicity of use, speed, and open source nature, MSpC combined with Morpheus is clearly advantageous over other PSM-based LFQ approaches currently available. Moreover, by being open source, MSpC should allow others to extend its utility and to serve as a platform for integrating additional open source LFQ approaches into the Morpheus pipeline.

## 2.3. Methods

## 2.4. Tutorial

## 2.5. ACKNOWLEDGEMENTS

D.C.G. was supported by a grant from the U.S. Department of Energy Office of Science; Office of Basic Energy Sciences; Chemical Sciences, Geosciences, and Biosciences Division (DE-FG02-88ER13968) and a graduate training fellowship from the NIH (5 T32 GM 7133-37). M.S. and L.M.S were supported by a grant from the National Institute of Health/National Institute of General Medical Sciences (1P50HG004942). The authors thank Erin Gemperline, Richard S. Marshall, and Josh Coon for critical reading of the manuscript, and additionally thank Derek Bailey for a critical code review.

## 2.6. References

**Book, A.J., Gladman, N.P., Lee, S.S., Scalf, M., Smith, L.M., and Vierstra, R.D.** (2010). Affinity purification of the arabidopsis 26s proteasome reveals a diverse array of plant proteolytic complexes. J. Biol. Chem. **285**(33):25554–25569.

**Cox, J., Hein, M.Y., Luber, C.A., Paron, I., Nagaraj, N., and Mann, M.** (2014). Accu-

rate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed maxlfq. Mol. Cell. Proteomics **13**(9):2513–2526.

**Deutsch, E.W., Mendoza, L., Shteynberg, D., Farrah, T., Lam, H., Tasman, N., Sun, Z., Nilsson, E., Pratt, B., Prazen, B., Eng, J.K., Martin, D.B., Nesvizhskii, A.I., and Aebersold, R.** (2010). A guided tour of the trans-proteomic pipeline. Proteomics **10**(6):1150–1159.

**Fermin, D., Basrur, V., Yocum, A.K., and Nesvizhskii, A.I.** (2011). Abacus: a computational tool for extracting and pre-processing spectral count data for label-free quantitative proteomic analysis. Proteomics **11**(7):1340–1345.

**Gerber, S.A., Rush, J., Stemman, O., Kirschner, M.W., and Gygi, S.P.** (2003). Absolute quantification of proteins and phosphoproteins from cell lysates by tandem ms. Proc. Natl. Acad. Sci. USA **100**(12):6940–6945.

**Ong, S.E., Blagoev, B., Kratchmarova, I., Kristensen, D.B., Steen, H., Pandey, A., and Mann, M.** (2002). Stable isotope labeling by amino acids in cell culture, silac, as a simple and accurate approach to expression proteomics. Mol. Cell. Proteomics **1**(5):376–386.

**Ross, P.L., Huang, Y.N., Marchese, J.N., Williamson, B., Parker, K., Hattan, S., Khainovski, N., Pillai, S., Dey, S., Daniels, S., Purkayastha, S., Juhasz, P., Martin, S., Bartlet-Jones, M., He, F., Jacobson, A., and Pappin, D.J.** (2004). Multiplexed protein quantitation in saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. Mol. Cell. Proteomics **3**(12):1154–1169.

**Russell, J.D., Scalf, M., Book, A.J., Ladror, D.T., Vierstra, R.D., Smith, L.M., and Coon, J.J.** (2013). Characterization and quantification of intact 26s proteasome proteins by real-time measurement of intrinsic fluorescence prior to top-down mass spectrometry. PLoS One **8**(3):e58157.

**Thompson, A., Schafer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., Neumann, T., Johnstone, R., Mohammed, A.K., and Hamon, C.** (2003). Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by ms/ms. Anal. Chem. **75**(8):1895–1904.

**Tu, C., Li, J., Sheng, Q., Zhang, M., and Qu, J.** (2014). Systematic assessment of survey scan and ms2-based abundance strategies for label-free quantitative proteomics using high-resolution ms data. J. Proteome Res. **13**(4):2069–2079.

**Wenger, C.D. and Coon, J.J.** (2013). A proteomics search algorithm specifically

designed for high-resolution tandem mass spectra. J. Proteome Res. **12**(3):1377–1386.

**Wong, J.W. and Cagney, G.** (2010). An overview of label-free quantitation methods in proteomics by mass spectrometry. Methods Mol. Biol. **604**:273–283.

**Zhang, B., VerBerkmoes, N.C., Langston, M.A., Uberbacher, E., Hettich, R.L., and Samatova, N.F.** (2006). Detecting differential and correlated protein expression in label-free shotgun proteomics. J. Proteome Res. **5**(11):2909–2918.

**Zhang, Y., Wen, Z., Washburn, M.P., and Florens, L.** (2010). Refinements to label free proteome quantitation: how to deal with peptides shared by multiple proteins. Anal. Chem. **82**(6):2272–2281.

**Zybailov, B., Mosley, A.L., Sardiu, M.E., Coleman, M.K., Florens, L., and Washburn, M.P.** (2006). Statistical analysis of membrane proteome expression changes in saccharomyces cerevisiae. J. Proteome Res. **5**(9):2339–2347.

# COLOPHON

This document was typeset with LaTeX . It is based on the University of Wisconsin dissertation template created by William C. Benton (available at `https://github.com/willb/wi-thesis-template`).